Opinion Spam Detection in Online Reviews

Ajay Rastogi[*] and Monica Mehrotra[†]
Department of Computer Science
Jamia Millia Islamia, New Delhi, India
[*]ajay148115@st.jmi.ac.in; ajayrastogijmi@gmail.com
[†]mmehrotra@jmi.ac.in; drmehrotra2000@gmail.com

Abstract: Online reviews are the most valuable sources of information about customer opinions and are considered the pillars on which the reputation of an organization is built. From a customer's perspective, review information is key to making a proper decision regarding an online purchase. Reviews are generally considered an unbiased opinion of an individual's personal experience with a product, but the underlying truth about these reviews tells a different story. Spammers exploit these review platforms illegally because of incentives involved in writing fake reviews, thereby trying to gain an advantage over competitors resulting in an explosive growth of opinion spamming. The present study analyzes and categorizes the available literature on opinion spamming according to three detection targets: (1) opinion spam, (2) opinion spammers, and (3) collusive opinion spammer groups. The study further highlights and divides opinion spamming into three types based on textual and linguistic, behavioral, and relational features. Moreover, several state-of-the-art machine-learning techniques for opinion spam detection have also been discussed in the study. It concludes with a summary of the research articles on opinion spam detection and some interesting results to assist researchers for further exploration of the domain.

Keywords: Opinion spam detection, Fake reviews, Opinion spamming features, Opinion spammer, Machine-learning techniques, Review trustworthiness

1. Introduction

In the recent years, the worldwide Web has radically changed the way people communicate and share their opinions globally. Online opinions are now expressed as posts, comments, reviews, or tweets on different online platforms like e-commerce sites, discussion forums, review sites, news sites, or any other social networking website. One of the ways of sharing an opinion is to write a review about a product or a service reflecting the customer's experience of that product or service. A customer believes in going through all the reviews about a product before deciding to purchase it. Therefore, these reviews are considered the basic unit of business and an eye-opener for business organizations and customers, respectively.

However, unfortunately, the question about the authenticity of these opinions has seldom received attention. "Are the reviews or opinions expressed by individuals authentic?" This remains a question mark for customers as well as business organizations. Product reviews contain highly valuable information about the product features and hence people are increasingly depending on them to make informed decisions before purchases (Jindal and Liu, 2008). This dependency has paved the way for fraudsters to gain incentives in exchange for putting up fake information to mislead consumers. Furthermore, since reviews can be freely written by anyone anonymously, there has been abuse of this anonymity feature by persons posting unethical or even fraudulent reviews to deceive consumers in order

to achieve significant business advantages (Liu, 2012). These reviews are written by people whose motives are highly suspect. Their sole aim is to endorse or downgrade the reputation of some target product or organization to gain personal profit or fame. Such reviews are called opinion spam or, more specifically, "fake reviews" (Jindal et al., 2010), and the persons indulging in these illegal activities are called "opinion spammers" (Mukherjee, Kumar et al., 2013). In this study, we have used the terms "opinion spam," "fake reviews," and "spam reviews" interchangeably. The increasing competition in the online marketplace has seen a growing trend of e-commerce companies hiring people to post fake positive reviews about their products or unfair negative reviews about their competitors' products in order to boost their own profit or to diminish the competitors' reputation (Xie et al., 2012a). Promoting products in this manner has become an established advertisement malpractice among online organizations. A number of cases1 have been reported in the past few years of the illegal marketing activities (such as hiring people to write fake reviews) of online business organizations.

Therefore, detection of opinion spam has become a big concern in present times to authenticate online opinions and gain consumer trust. Jindal and Liu (2007b) pioneered the research on the problem of opinion spam by analyzing a large dataset of Amazon product reviews. The authors provided a basic categorization of fake reviews: untruthful reviews, reviews on brand only, and non-reviews, along with an easy solution to detect them. With their research as a base, a number of other researchers have tried to explore the different features of opinion spam and methods to solve the problem effectively and efficiently. In the literature, researchers have viewed opinion spam detection from three different angles: detecting fake online reviews (i.e., opinion spam) (Jindal and Liu, 2007a; Jindal and Liu, 2007b; Jindal and Liu, 2008; Lau et al., 2011; Xie et al., 2012a; Li et al., 2013), detecting persons involved in writing fake reviews (i.e., opinion spammers) (Lim et al., 2010; Wang et al., 2012; Fei et al., 2013; Mukherjee, Kumar et al., 2013), and detecting networks of opinion spammers (i.e., opinion spammers' groups) (Mukherjee et al., 2011; Mukherjee et al., 2012; Xu et al., 2013; Xu and Zhang, 2015b; Ye and Akoglu, 2015). A variety of effective spamming features such as review-content-specific features (Ott et al., 2011; Feng, Banerjee et al., 2012; Ong et al., 2014), reviewer-behavior-specific features (Lim et al., 2010; Feng, Xing et al., 2012; Xie et al., 2012a), and review-reviewer network-specific features (Wang et al., 2012; Akoglu et al., 2013; Rayana and Akoglu, 2015) have been used in studies to successfully detect spam reviews. The most promising methods used by several researchers include machine-learning techniques (Guzella and Caminhas, 2009) that have offered some interesting results to aid fake review detection. From the beginning of research on opinion spam, researchers have employed supervised methods (Ott et al., 2011; Li et al., 2013; Banerjee et al., 2015) to train their classifiers. However, due to the lack of ground-truth datasets, they later resorted to using unsupervised methods (Lau et al., 2011; Wang et al., 2012; Mukherjee, Kumar et al., 2013) to identify spam reviews. Some works have effectively used semi-supervised methods, for example, the co-training framework (Li et al., 2011) and positive-unlabeled (PU) learning framework (Fusilier et al., 2014; Li, Chen et al., 2014), to overcome the problem of fake review annotation.

[1] http://www.nytimes.com/2012/01/27/technology/for-2-a-star-a-retailer-gets-5-star-reviews.html?_r=3&ref=business
http://www.thedenverchannel.com/news/woman-paid-to-post-five-star-google-feedback
http://www.businesstoday.in/technology/news/amazon-sues-1000-people-for-fake-online-iphone-product-customer-reviews/story/225069.html

Another category of methods to identify opinion spam includes text mining and Natural Language Processing (NLP) techniques. By considering the review's text as a major ingredient of opinion spamming, researchers believe that opinion spammers always leave behind linguistic clues in deceptive text writing. Language modeling techniques (Lai, Xu, Lau, Li and Jing, 2010; Lau et al., 2011; Li et al., 2013), N-grams methods (Ott et al., 2011; Ott et al., 2013), Part-Of-Speech (POS) tagging information (Feng, Banerjee et al., 2012; Li, Ott et al., 2014), duplicity measure (e.g. cosine similarity) (Jindal and Liu, 2008), Term frequency-Inverse document frequency (Tf-Idf) measure (Al-Najada and Zhu, 2014), and so on have been extensively utilized in the literature to detect deception in a review's text component.

In this paper, we have provided an in-depth review on the problem of opinion spam detection with a brief summary of the research articles published between 2007 and 2015. To gain a better understanding of the different aspects of opinion spam detection, the literature has been categorized based on three factors: detection targets, spamming features, and opinion-spam detection methods. The first category discusses the different detection targets of opinion spam, which include detecting opinion spam, opinion spammers, and collusive opinion spammers' groups. Thereafter, the different spamming features referred to in previous works have been grouped into three categories based on textual and linguistic features, behavioral features, and relational features. The third category includes a variety of opinion-spam detection methods with a focus on their results. Furthermore, a brief analysis has been presented summarizing the different views on opinion spamming. This study can be of immense benefit to new researchers and practitioners in offering in-depth knowledge on opinion spam detection and raises the issue of important research gaps for further improvement and exploration.

One of the main reasons for conducting this comprehensive literature review is the lack of review articles on opinion spam detection. Although there are a number of studies on detection of fake reviews, to the best of our knowledge, only a few articles have summarized the research work on this topic. For example, Sheibani (2012) summarizes only opinion spam and opinion mining at the very basic level; Heydari et al. (2015) illustrate the various methods used to detect opinion spam, opinion spammers, or their groups; and Crawford et al. (2015) explain the various features and machine-learning techniques employed by previous studies to spot fake opinions. All these articles present different views about opinion spam, but none of them provides a complete review of the problem. In this study, we attempt to cover almost all aspects of opinion spam detection, provide an overview of the reputed research articles published between period 2007 and 2015, and organize them in a systematic review to project the current progress in the domain.

The section-wise division of this study is as follows: Section 2 introduces the problem of opinion spam along with a basic categorization of the different types of spam. Data collection methodology is discussed in section 3. Section 4 categorizes the literature according to the three different detection targets. Another method of categorizing spamming features used to detect opinion spam is presented in section 5. In section 6, various machine-learning techniques employed by previous works have been discussed along with their results. A brief summary and some useful results of our survey are listed in section 7. Section 8 offers a conclusion of this work.

2. Opinion Spam: An Overview

Opinions are the central part of any post, comment, tweet, or review. Spam refers to any irrelevant or unsolicited information attached with these opinions for the purpose of advertisement, promotion,

information spread, or even financial profit (Delany et al., 2012). Spam is not a new field of research; a number of other types of spam such as e-mail spam (Guzella and Caminhas, 2009), web page spam (Spirin abd Han, 2012), social spam (Chakraborty et al., 2016), sms spam (Delany et al., 2012), and so on has already been examined extensively in the literature on spam filtering. A new inclination in spam-detection research is the investigation of efficient opinion spam filters, which can be effective in identifying opinion spam more accurately. This paper focuses on opinion spam and its detection in detail.

"Opinion spamming" refers to the provision of wrong or false information in reviews to misguide consumers and influence product sales (Jindal and Liu, 2008). Product reviews are generally used by individual customers to make online purchase decisions, that is, whether to buy a particular product or not. Positive reviews attract customers and vice versa (Xie et al., 2012a). This reliance of consumers on reviews has led e-commerce organizations to making wrong claims or even indulging in illegal activities such as hiring persons to write fake positive reviews in their favor and unfair negative reviews about their competitors. Persons (opinion spammers) who provide a fake review are paid money or given free coupon codes in exchange. Therefore, filtering these fake reviews is of utmost importance in retaining authenticity of online reviews as well as consumer trust in e shopping.

2.1. Categorization of Opinion Spam

Jindal and Liu (2007b) categorize opinion spam into false or untruthful opinions, reviews on brands only, and non-reviews. They found that the first type, spam reviews (untruthful reviews), are more damaging than the other two types, and comparatively more difficult to detect. Untruthful opinions can further be divided into positive opinion spam (Hyper Spam/Ballot Stuffing) (Ott et al., 2011) and negative opinion spam (Defaming Spam/Bad Mouthing) (Ott et al., 2013). Hyper spam includes fake reviews exhibiting positive sentiment and written by suspicious persons unjustly promoting the reputation of some desired organization (or product), while defaming spam refers to fake reviews showing a competing organization (or product) in bad light by writing something defamatory about it or its products. Harris (2012) considers untruthful opinions as deceptive opinion spam, while the other two as not deceptive (i.e., disruptive opinion spam). Different types of spams are shown in Fig. 1 further categorizing opinion spam into two general categories: Deceptive opinion spam and Disruptive (but not deceptive) opinion spam (Ott et al., 2011). Generally, opinions or reviews containing only advertisements (like URLs or self-endorsements) or random texts are not considered as deceptive spams, and are much easier to detect by manual inspection. In the current literature on opinion spam detection, the maximum number of studies is centered on detection of deceptive opinion spam, which is more harmful and much harder to detect as compared to disruptive opinion spam.

2.2. Structure of Online Reviews

Online opinions can be expressed in various forms such as in a comment, post, status, tweet, or review. In this paper, we limit our discussion to opinions typically expressed in the product or store reviews posted on various e-commerce sites. Typically, a product review is made up of the product's unique identification (ID) number, the reviewer's (user's) unique identification (ID) number, the number of helpfulness votes the review received, the rating on the 5-star rating scale, the review body containing the reviewer's opinion, the review title summarizing the review, the verified purchase tag showing that the user has actually purchased the product, and the review date and posting time as shown in Fig. 2. These are the external visible properties of reviews that were used by different studies in the literature to

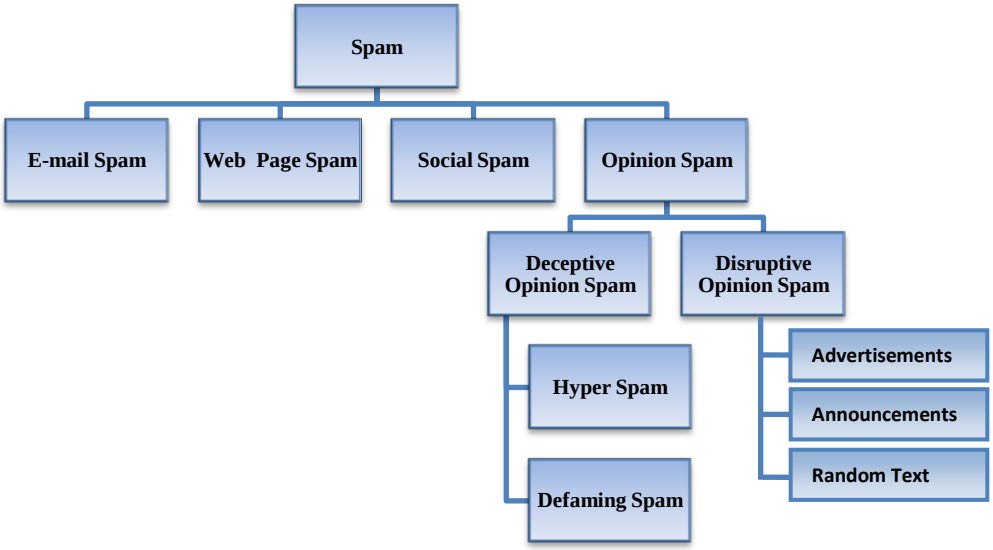

Fig. 1. Categorization of opinion spam.

distinguish spam reviews from non-spam ones. Besides these, a review also has a few internal parts that contain private information not displayed on the review sites, such as the IP address or MAC address of the reviewer. We will further discuss these review features used to detect opinion spam in section 5.

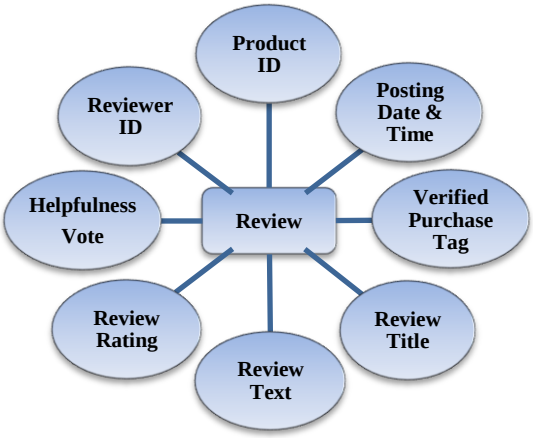

Fig. 2. Basic components of an online review.

3. Methodology

In this section, we describe the data collection process to collect the literature on opinion spam detection, as shown in Fig. 3. We first searched for the term “opinion spam detection” from the Web search engine “Google Scholar” and found some of the published research articles closely related to our search query and considered them as the base papers for further searches. After carefully scrutinizing these base papers and gaining some insights into opinion spamming, we looked for similar terms related to opinion spam, such as fake reviews, bogus reviews, deceptive reviews, untruthful reviews, review spam, forged reviews, shill reviews, review spammers, opinion spammers, and so on. Then, we searched for different combinations of these terms on different online platforms, such as IEEE Explore digital

library, ACM digital library, Elsevier's ScienceDirect, Google Scholar, ResearchGate network, SpringerOpen, and so on and came across a number of journal- and conference-level publications. Alongside, we filtered out the irrelevant or poor-quality research articles from our collected literature by applying quality measures like indexing (e.g. sci indexed, scopus indexed, indexed in Thomson Reuters, scimago journal ranks etc.) for journal-level publications and ranking/grading (e.g., A, A+, B, B+ ranking conferences) for conference-level publications along with their publishers (e.g., IEEE, ACM, Springer, AAAI, Elsevier etc.).

To maintain the quality of our reviewed literature, we retained only those articles published by well-known publishers in high-indexed journals having a good impact factor or presented at top-level conferences. We also applied different filtering mechanisms like a year-wise search or sorted by date or relevance from the Web search engines/digital libraries to enhance the quality of our collection process.

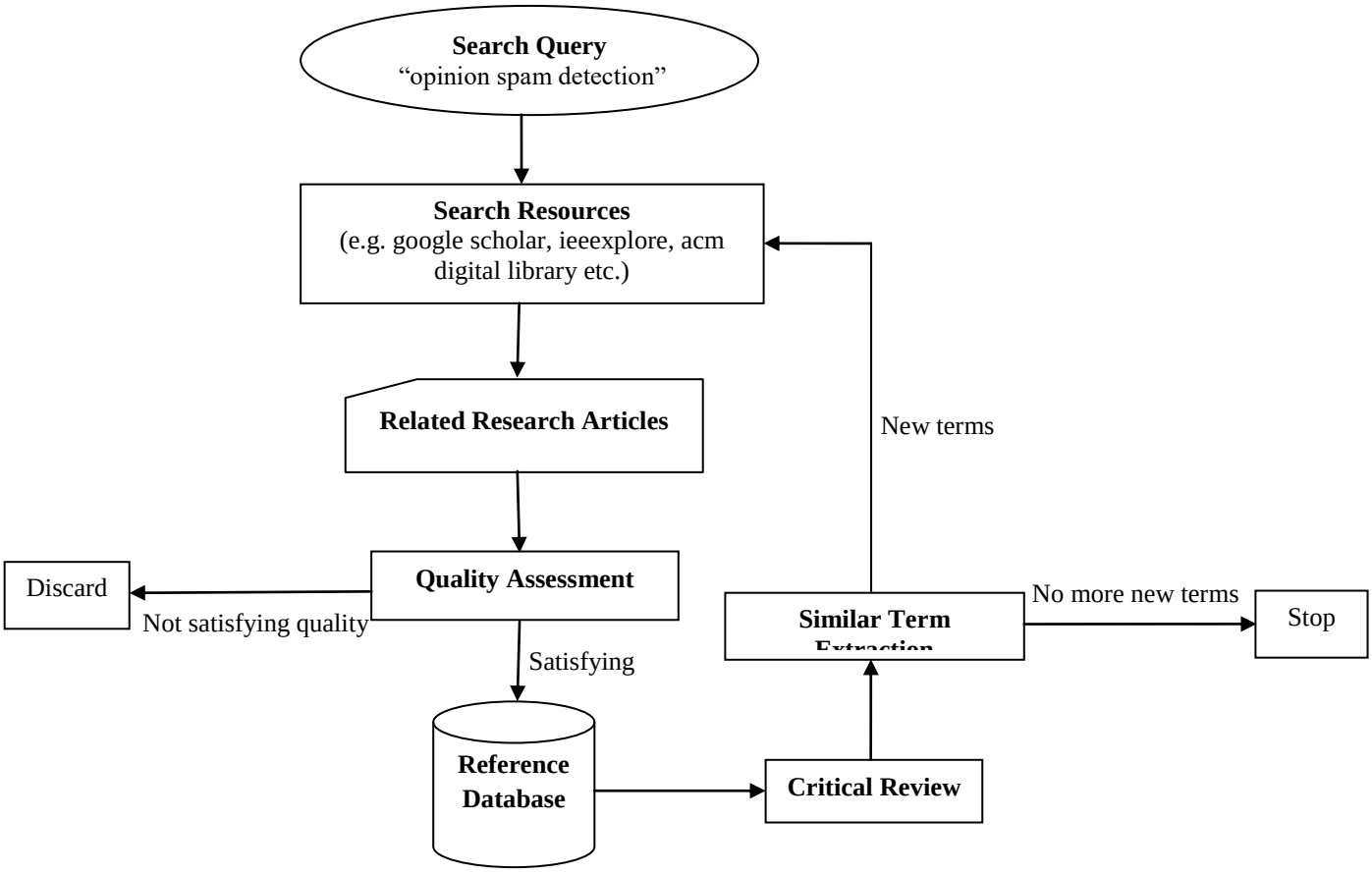

Fig. 3. Data collection methodology.

In Figure 3, our schema consists of several steps to collect the literature concerning "opinion spam". The first step was "search query" which contains some common terms related to opinion spam detection. Then, we applied search process by using different search engines and retrieved some research papers. In the next step, papers were filtered based on presumed quality criteria. Afterwards, we carefully reviewed the collected papers and found more terms related to "opinion spam". Then again, search was performed

based on newly collected terms and more relevant literature was added in reference database. This process was iterated until no more new terms were found during the search.

Furthermore, to obtain a sufficient amount of literature on opinion spam detection, we also examined the references in our collected research articles and retained the best ones related to our topic. Our collection includes almost all the relevant research articles on fake review detection published in the period between 2007 and 2015 organized in a systematic review to understand the current research progress so far.

4. Categorization Based on Opinion-Spam Detection Targets

Since opinion spam exists in a variety of forms, such as spam posts, spam comments, spam status, spam blogs, spam reviews, and so on, each of these forms requires a different detection mechanism according to its distinct nature. Therefore, collecting, reviewing, and categorizing all the literature covering the different types of opinion spam is a difficult task. For example, in online discussion forums, opinion spam results in the creation of new threads to divert the discussion topic into some intended direction (Chen and Chen, 2015); on micro blogging platforms, the target of spammers is to influence a particular post or blog (Lee et al., 2015); on news websites, the spammers can participate in any event to post spam social comments (Chen et al., 2015). This paper discusses the opinion spam existing specifically in online product or store reviews. Moreover, among all types of opinion spam, review spam detection has been given maximum attention by researchers in the last decade. Researchers have focused on different detection targets to identify deception in reviews. In this section, the existing works on opinion spam detection have been summarized according to three targets: detection of (1) opinion spam, (2) opinion spammers, and (3) collusive Opinion spammers' groups.

4.1. Detection of Opinion Spam

One of the most explored dimensions in the field of opinion spam is the detection of fraudulent reviews written by persons whose intentions are suspect. In the early stage of opinion-spam detection research, scientists focused on detection of opinion spam or fake reviews. Later, other detection targets, such as opinion spammers and their groups, were investigated to ultimately detect the identity of persons involved in opinion spamming. Although quality assessment of product reviews is necessary to help customers in decision making, mere quality cannot be a spam indicator. Liu et al. (2007) discriminate between low- and high-quality reviews based on the textual information provided by reviewers, but do not indicate a spam attack in the reviews. A low-quality review can be the output of a genuine user who may not have enough experience to write a sound review. While, on the contrary, experienced fraudsters can write high-quality reviews (Liu et al., 2007). Therefore, it becomes necessary to analyze factors other than the quality of a review to detect spam reviews. The first attempt to detect opinion spam was made in the year 2007 (Jindal and Liu, 2007a; Jindal and Liu, 2007b) in which a massive Amazon review dataset was analyzed to indicate some spamming clues in reviews, and a possible categorization of opinion spam was presented as untruthful reviews or false opinions, reviews on brands only, and non-reviews. Studies indicate that to earn profit, opinion spammers try to copy their own or other customers' reviews and post them with a little manipulation (Lau et al., 2011). This allows them to save time by not having to write a new review each time (Mukherjee, Kumar et al., 2013). Considering opinion spam as a binary classification problem, duplication-detection techniques were employed (Jindal and Liu, 2007b; Jindal and Liu, 2008) in which text similarity was the main evidence to spot fake opinions. The reviews having

maximum similarities with other reviews about the same product (or posted by the same user) were fall in the category of spam and not as duplicated reviews, which had some legitimacy. Algur et al. (2010) suggest that instead of spotting the duplication in a review's text, one can measure the duplication in a review's concept (product features mentioned in reviews) to identify an opinion as fake or legitimate. They classify the reviews into four types: exact duplicate, near duplicate, partial duplicate, and unique, in which the former two types were treated as spam reviews and the latter two as genuine. Moreover, by gaining some important clues from information retrieval, some studies proposed language modeling approaches such as probabilistic language modeling with KL-divergence and Support Vector Machine (SVM) classifier (Lai, Xu, Lau, Li and Jing, 2010), text mining with content overlapping (Lau et al., 2011), language modeling with concept association (Lai, Xu, Lau, Li and Song, 2010), and so on to identify untruthful reviews by analyzing the reviews' syntactic and textual features. Lau et al. (2011) adopt a different duplicity measure called semantic content overlapping and use a unigram language model with KL-divergence to spot fake reviews. The authors show a 97% true positive rate of opinion spam detection, which is comparatively better than the rate in other studies.

One of the major challenges in the detection of fake reviews is the unavailability of ground-truth datasets for model training (Ma and Li, 2012), and the situation is worse in the case of an imbalance in the data of fake and truthful reviews (Al-Najada and Zhu, 2014). As a result, several studies have employed a manual labeling process to evaluate their classifier's performance. In this regard, Ott et al. (2011) created a ground-truth labeled dataset of 800 reviews, which includes both truthful and deceptive reviews, for model training. According to their findings, deceptive reviews are generally "more imaginative" while truthful reviews are "more informative" in nature. In this way, they provide a clear distinction between the two classes of reviews, that is, deceptive and truthful reviews. Mukherjee, Venkataraman et al. (2013a) show that many existing studies (Harris, 2012; Li et al., 2013; Li, Ott et al., 2014) have used pseudo fake reviews generated through crowdsourcing tools like Amazon Mechanical Turk (AMT) to train their classifiers and report greater accuracies (around 90% (Ott et al., 2011), 91.2% (Feng, Banerjee et al., 2012), 94.8% (Li et al., 2013)). Pseudo fake reviews are produced by the persons hired to generate fake reviews for model training purposes. The authors find that these pseudo fake reviews are not good representatives of actual fake reviews, and that detection algorithms tested on pseudo fake reviews lead to low accuracies when applied on actual fake reviews.

Although Ott et al. (2011) produced a gold-standard dataset of fake reviews, it took them a good amount of cost and time to create and label each review manually. To overcome this problem, Feng, Xing et al. (2012) proposed an approach to automatically label the reviews based on the distributional clues left by the raters. This study suggests that the average rating of malicious reviewers (especially single-time reviewers) differs significantly from those of truthful reviewers for a fraudulent hotel or store. Accordingly, they proposed an alternative evaluation approach that does not rely on manual human labeling. Again, to suppress the cost involved in hiring human turkers to write fake reviews, Sun et al. (2013) proposed a synthesis process to generate fake reviews that closely resemble truthful reviews. They suggest that fake reviews can be automatically generated by replacing the sentences of a truthful review with the sentences of some other reviews in the review corpus. Furthermore, most of the existing detection algorithms fail to detect such synthetically generated fake reviews. In order to detect these synthesized fake reviews, they proposed using sentence-wise similarity measures and reported greater accuracies (approx. 13 % improvement) over the existing approaches. Some of publicly available ground-truth datasets that have been employed by most of the works listed in Table 1. The maximum number of

Table 1. Some publicly available datasets.

Dataset Name	Description	Ground Truths	Download URL	References
Amazon Product Review Dataset	5.8 million reviews, 2.14 million reviewers and 6.7 million products.	Not Available	Amazon dataset url[2]	(Jindal and Liu, 2007a; Jindal and Liu, 2007b; Jindal and Liu, 2008; Jindal et al., 2010; Lim et al., 2010; Mukherjee et al., 2011; Feng, Xing et al., 2012; Lappas, 2012; Mukherjee et al., 2012; Fei et al., 2013; Lu et al., 2013; Mukherjee, Kumar et al., 2013; Liang et al., 2014; Lin, Zhu, Wang et al., 2014; Lin, Zhu, Wu et al., 2014; Savage et al., 2015; Wang, Hou et al., 2015; Xu and Zhang, 2015b)
TripAdvisor Hotel Review Dataset	400 positive truthful reviews from TripAdvisor, 400 positive fake reviews from AMT. 400 negative truthful reviews from TripAdvisor, 400 negative fake reviews from AMT.	Available	TripAdvisor dataset url[3]	(Ott et al., 2011; Feng, Banerjee et al., 2012; Li et al., 2013; Montes-y-Gómez and Rosso, 2013; Mukherjee, Venkataraman et al., 2013a; Ott et al., 2013; Al-Najada and Zhu, 2014; Fusilier et al., 2014; Li, Ott et al., 2014; Fusilier et al., 2015; Karami and Zhou, 2015; Sandulescu and Ester, 2015)
Yelp Filtered Review Dataset	Three Versions: YelpCHI – 67,395 reviews, 38,063 reviewers, 201 hotels and restaurants from Chicago. YelpNYC – 359,052 reviews, 160, 225 reviewers, 923 restaurants from New York. YelpZip – 608,598 reviews, 260,277 reviewers, 5,044 restaurants.	Available	Yelp dataset url[4]	(Mukherjee, Venkataraman et al., 2013a; Mukherjee, Venkataraman et al., 2013b; Rayana and Akoglu, 2015)
Dianping Restaurant Review Dataset	9,765 reviews, 9,067 users, 5,535 IPs, 500 restaurants from Shanghai, China.	Available	Dianping dataset url[5]	(Li, Chen et al., 2014; Li et al., 2015)

[2] http://liu.cs.uic.edu/download/data/
[3] http://www.cs.cornell.edu/~myleott/op_spam/
[4] http://liu.cs.uic.edu/download/yelp_filter/
http://odds.cs.stonybrook.edu/yelpchi-dataset/
http://odds.cs.stonybrook.edu/yelpnyc-dataset/
http://odds.cs.stonybrook.edu/yelpzip-dataset/
[5] http://liu.cs.uic.edu/download/dianping/

papers in the literature adopted either the Amazon or the TripAdvisor dataset for their model evaluation. However, building a scalable and domain-independent ground-truth dataset for opinion spam detection is still an open challenge for researchers.

A new dimension in opinion-spam detection research is to identify a Singleton Review (SR) spam (Xie et al., 2012b). If a reviewer writes only one review from an account, then the review is termed as a "singleton review." It is easy to catch the behavior of a spammer who writes multiple reviews, but very difficult to spot a spammer who has written only one or two reviews, which is normally the case. Xie et al. (2012a) scrutinized such SR spam and indicated that SR spam usually come in bursts and distort the average rating significantly. The authors proposed a temporal approach and constructed a multi-scale multi-dimensional time series to find the correlation between singleton review arrival pattern and average ratings given to a store and detected those time windows in which the average ratings of the store increase sharply with the arrival of singleton review bursts to highlight SR spam. Similar work by Wu et al. (2010) suggests how positive singletons appear in quick succession and manipulate the true rankings assigned to a hotel. Most of the studies have simply discarded this view of review spam by neglecting reviewers who have written only one review, but this aspect is equally important since a large portion of online review platforms is covered by singleton reviews. Therefore, methods more effective need to be explored in order to protect review platforms from singleton review spam attacks.

Moreover, studies suggest that if a reviewer is writing multiple reviews within a short time interval (e.g. within a week), it clearly reflects a suspicious case. Real users normally write very few reviews in a short time, while spammers write as many reviews as possible over a short period to boost their profit. Some time-sensitive behavioral features (Lin, Zhu, Wang et al., 2014; Lin, Zhu, Wu et al., 2014) of reviewers can be applied to detect such fake reviews written by a bursty reviewer. Hu et al. (2011) suggest a temporal approach to determine the relationship between average product ratings and review manipulation. They showed that average consumer ratings toward a product decrease with the passage of time. This clearly indicates that early reviewers have a great impact on product ratings and stand a greater chance of indulging in review manipulation. Online reviews should only express an opinion on the target product, but often, some customers have been found to give high ratings along with a negative comment in the review text or vice versa. To detect such suspicious opinions, Sharma and Lin (2013) devised a tool that takes the review rating and text as an input and indicates the inconsistency between review sentiment and the rating given by the reviewer.

A deceptive review can be positive or negative as indicated in section 2. In the previous literature, a large proportion of studies applied various machine-learning techniques to detect deceptive positive spam reviews (Ott et al., 2011; Fusilier et al., 2014; Li, Ott et al., 2014), while only a few focused on detecting negative spam reviews (Ott et al., 2013). As negative spam reviews are primarily responsible for diminishing the reputation of an organization, these are equally harmful. This has led to a research gap, and future research needs to focus on this neglected part of opinion spam detection.

4.2. Detection of Opinion Spammers

The next promising dimension is detection of opinion spammers who are responsible for posting malicious reviews to deceive readers. Researchers believe that it is easier to spot a spammer than a fake review because a review provides limited information while an opinion spammer provides several clues by writing multiple spam reviews. Earlier literature was diverted toward opinion spam detection (review-

centric), while recent studies have targeted opinion spammers (user-centric), who are usually hired by organizations to gain unfair profits (Lim et al., 2010; Sandulescu and Ester, 2015). Spammers may target specific products or stores to achieve their goals. Since some websites provide product reviews (like Amazon.com) while some others offer store reviews (ResellerRatings.com), researchers have focused on two different types of spammers and suggest using different methods to detect them. In the subsequent subsection, we discuss two types of review spammers and their detection techniques: product-review spammer detection and store-review spammer detection.

4.2.1. *Product Review Spammer Detection*

A product can be anything ranging from a mobile device, movie, or restaurant to an online service, and product reviews characterize the features of the product along with their star ratings. Most of the previous studies have focused on review text information to mark a review as fake or non-fake, but to identify opinion spammers, it is necessary to trace spamming behavior clues of reviewers (Mukherjee, Kumar et al., 2013). For example, if a reviewer writes multiple reviews on the same product or on the same brand of products by marking only high or low ratings and/or repeating text each time, then there is clearly room for suspicion. Lim et al. (2010) observed that spammers could target specific products or product groups (e.g., different products of the same brand) and deviate from normal reviewers in terms of their rating patterns as well as the review text information they provide. Therefore, they proposed two models of spamming behaviors, target-based spamming and deviation-based spamming, to label the reviewers as spammers or non-spammers. The authors assign a spamicity score to each reviewer by combining the two models and treat the reviewers having high spamicitiy score as more likely to be spammers. In order to lessen the cost involved in the analysis of review text, Savage et al. (2015) proposed a pure-rating-based approach using binomial regression to model reviewers' rating behavior and spot reviewers with a large proportion of reviews deviating from the mean ratings as highly suspicious candidates for review spamming. Again, to highlight the spurious behavior of a spammer, Jindal et al. (2010) employed Class Association Rule (CAR) mining to form interesting and unexpected rules showing the anomalous rating patterns of spammers. More specifically, their approach was domain independent and therefore applicable to other domains as well. A novel rating-based method was provided by Allahbakhsh and Ignjatovic (2015), in which the effects of fake ratings are reduced in calculating the final rating for a product under consideration. By considering the posted time as a crucial factor, Fei et al. (2013) demonstrated the temporal behavior of spammers who post reviews in small time bursts. They employed Kernal Density Estimation (KDE) function to indicate reviews bursts, modeled the reviewers in bursts as Markov Random Fields (MRF), and adopted a Loopy Belief Propagation (LBP) algorithm to identify the likely networks of spammers in review bursts. Along with the temporal features, Li et al. (2015) used some spatial features (IPs and city locations) of reviewer behavior to distinguish product review spammers from non-spammers in the restaurant domain. The authors hypothesized that expert spammers frequently switch IP addresses to post multiple reviews. By combining the two features, they indicate some spamming clues like frequent and random movement between cities by a user when posting reviews and frequent switching of IPs by the same user. However, IPs are generally not available publicly, and therefore, this type of research is not always possible.

Furthermore, since most of the studies have focused on detection of spam reviews or review spammers separately, it would be more appropriate to detect both opinion spammers and spam on a unified framework. By observing the correlations among reviews, reviewers, and products, some graphical

frameworks were proposed in the literature (e.g., FRAUDEAGLE (Akoglu et al., 2013), Review Factor Graph (RFG) model (Lu et al., 2013), SPEAGLE (Rayana and Akoglu, 2015)) to collectively detect fake reviews and review spammers. Akolgu et al. (2013) created a bipartite network of reviewers and products associated with review sentiment and labeled each reviewer as fraudulent or honest, each product as good or bad, and each review as fake or real by propagating the label information between different nodes of the network. Figure 4 illustrates this association among reviews, reviewers, and products on a tripartite graph. A drawback of FRAUDEAGLE is that it cannot incorporate additional prior information in the network on its own, such as review text, behavioral clues, and so on. To overcome this problem, another framework, SPEAGLE (Rayana and Akoglu, 2015), was proposed, which could easily utilize the metadata information of reviews such as its text, ratings, timestamps, and so on in the user-review network and enhance the performance of FRAUDEAGLE considerably.

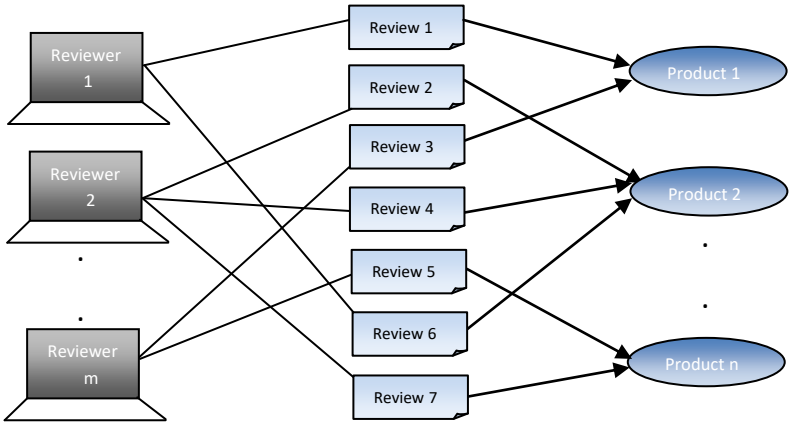

Fig. 4. Tripartite graph representing the association among reviews, reviewers, and products.

Similar to Xie et al. (2012a), another effort was made to spot singleton review spammers (Sandulescu and Ester, 2015). The authors suggest that spammers use multiple IDs or names to cast their reviews and hence it is difficult to identify them by simple spamming filters. To detect the reviews of a reviewer written under different names, two methods, called semantic similarity measure and topic-modeling-based similarity measure, were proposed. The authors found that reviews written using different user-IDs on a product and in a semantically similar context (satisfying a certain threshold) were the most suspicious ones and probably written by the same person using slight modifications. They demonstrated the role of a singleton review spammer who writes only a single review or few reviews and uses multiple accounts to cast each review in order to avoid detection by simple spamming filters.

4.2.2. *Store Review Spammer Detection*

Another dominant area in opinion-spam detection research is the detection of store review spammers. The difficulty in understanding the behavior of store review spammers has resulted in only a few attempts to identify such spammers. Store review spammers are different from product review spammers as their aim is to praise or criticize the reputation of a target store by providing unfair ratings and reviews about a whole store rather than on the specific products of the store. Resellerratings.com, Bizrate.com etc. are the main platforms to write store reviews.

Wang et al. (2011) and Wang et al. (2012) were the first to spot with acceptable accuracies store review spammers. Wang et al. (2011) claimed that existing behavior-based algorithms are helpful in detecting product review spammers but are inadequate to catch the abnormal behavior of store review spammers, and therefore they proposed a new review-graph-based approach by incorporating the association among honest reviews, trustworthy reviewers, and reliable stores. First, they define the three prominent concepts of honesty of reviews, trustworthiness of reviewers, and reliability of stores with observations as shown below:

i) A review's honesty directly depends on the reliability of the store to which it belongs and its conformity with other reviews of the same store.
ii) A reviewer's trustworthiness is directly related to the number of honest reviews written by that reviewer.
iii) A store's reliability is directly related to the number of positive reviews written by the most trustful reviewers.

As all the three features described above can only be computed with the help of each parameter, a heterogeneous tri-partite graph was created to handle the inter-dependencies among them. Finally, an iterative computational model was employed to find the most distrustful reviewers and they were labeled as spammers. A similar reviewer-store graph-based approach was employed by Peng (2013) in which the authors tried to identify store review spammers who use single user-IDs or multiple IDs to cast spam opinions. Authors defined these two types of spammers as single-mode spammers and multi-mode spammers, respectively. Their work presents two important clues to spot both types of spammers: a semantic deviation of a review from the overall community sentiment for a store and a rating deviation from the mean ratings given to a store. As online shopping stores are the primary source of opinion spamming, it would be better to identify suspicious stores responsible for posting fake reviews rather than identifying fake reviews or review spammers. Once a store is blacklisted, all reviews of that store can be considered fake and customer trust regarding that store could change accordingly. The latest work of Chengzhang and Kang (2015) identified these suspicious stores by concentrating on the skeptical behavior of spamming stores in terms of targeted deals and reviews. No other work in the literature has been found to support this view of opinion spamming, thereby giving rise to a research gap.

Identification of opinion spammers has attracted the attention of researchers since opinion spammers are the main source of spamming, but due to the ever-changing behavior of spammers, existing detection techniques will not work for a long time. As most of these detection techniques use fixed parameters, spammers may learn to change their behavior accordingly and remain successful in spamming. Therefore, researchers need to develop effective spamming filters that can capture the dynamic behavior of opinion spammers to prevent their violations in future.

4.3. Detection of Collusive Opinion Spammers' Groups

A few recent studies have highlighted a new dimension to detect suspicious opinion spam cases, which we call "detection of collusive opinion spammers' groups." As spammers are hired for monetary purposes, to gain more profit, and to create a greater impact of their opinions, they are likely to work in collaboration. A spammer group is a group of people working collectively by posting deceptive reviews to multiple target entities, and these groups are more hazardous than individual spammers since as a group they hide behind the true sentiment of genuine reviewers' opinions and thereby largely nullify the

trustworthiness of opinion sharing websites (Mukherjee et al., 2011). Mukherjee et al. (2012) were the first to introduce the concept of group spamming in the context of product reviews. First, they analyzed a huge Amazon dataset and applied Frequent Itemset Mining (FIM) to find candidate groups of reviewers who have reviewed at least three (support count) products collectively. They then proposed the following spamming features to spot spam groups of reviewers: reviewers of a group posting reviews over a small period; average ratings of group members different from that of other members who are not part of the group; review text similarity of reviewers within the group; text similarity of individual reviewers of the group; number of reviewers in the group; and so on. In addition, they also built a labeled dataset of spam and non-spam groups by using human judgment and applied a relation-based approach GSRank to rank the identified groups according to their likelihood of being spam. After experimental evaluation, they showed that the algorithm GSRank achieves greater success over the standard supervised classification approaches like SVM and Regression. The latest work of Xu and Zhang (2015b) argues that features provided by Mukherjee et al. (2012) are not concrete enough to grab the more elusive spamming behavior of collusive spammers. They proposed other interesting measures like consistency in targeted businesses, rating consistency, temporal synchronization, first-review synchronization, activity similarity, and workload similarity to identify narrow collusive spamming groups. Moreover, studies suggest that it is easier to detect colluders than individual fake reviews or review spammers as they leave behind more clues of their collaborative abnormal behavior in the targeted network. Since spammers create a kind of network to violate reviews, a majority of studies applied graph-based approaches (Choo et al., 2015; Wang, Hou et al., 2015; Ye and Akoglu, 2015) to spot these groups. Spammers often distort the statistical properties of a real-world network. Therefore, using a product-review bipartite graph (Wang, Hou et al., 2015) and some network-related structural properties (e.g., centrality measures) (Ye and Akoglu, 2015) can help in the easy detection of these spammers from the highly suspicious, abnormal patterns of the graphs.

Moreover, existing methods are able to detect fake reviews written in English. A report[6] indicates that China's internet users are also exceedingly becoming adept in providing fake reviews on various review websites. To tackle the problem of opinion spamming in Chinese websites, some works (Xu, 2013; Xu et al., 2013) analyzed Chinese-language datasets. Findings suggest that most of the existing techniques fail to seize the collusive spamming behavior of Chinese reviewers. In order to boost the performance of the traditional classification approach, Xu (2013) proposed a hybrid approach of classification into clustering tasks to form groups of colluders by utilizing their relational characteristics. However, overall, the existing state-of-the-art approaches are domain- and language dependent. Therefore, it is of utmost priority to develop such hybrid techniques that can be applied in any review domain effortlessly. In addition, only a few studies have used the structural properties of the reviewer network to identify suspicious spamming groups. Since spammers may find it difficult to distort the structural parameters of a network, there is a need to explore more sophisticated network parameters to detect complex colluding groups of reviewers.

5. Categorization Based on Spamming Features

[6] http://www.independent.co.uk/news/business/news/amazon-fake-reviews-came-from-bangladesh-china-the-philippines-and-the-uk- investigation-finds-a6699956.html

Feature extraction is always considered a crucial step in the process of identifying opinion spam. The effectiveness and accuracy of any spam-filtering algorithm depends on the input features provided to that algorithm. The more relevant the input features, the more valid the output in an algorithm. A major challenging part in opinion spam detection is the identification of optimal features that best describe the spamming behavior of malicious reviewers. Since a review comprises three main components, review's text, reviewer, and product, as described in section 2, studies suggest using review-text-related features, reviewer-related features, and product-related features (Jindal and Liu, 2007b; Jindal and Liu, 2008; Li et al., 2011). After carefully analyzing different approaches to detect opinion spam, we identified three sets of features: linguistic and textual features, behavioral features, and relational features.

5.1. Linguistic and Textual Features

The language that spammers use reflects their feelings, thoughts, and emotions. When someone writes a deceptive text, some clues of linguistic deception can be seen in the language used. Linguistic features are one of the key indicators to detect opinion spam and help in summarizing reviewers' text in the form of a vector. Opinion spammers can have a different writing style than that of non-spammers (Ott et al., 2011), or they may reproduce a spam review by copying their own reviews; therefore, it is better to focus on their language and writing styles to detect malicious reviews. A number of studies have been conducted to capture the linguistic clues left behind by spammers. Some of the commonly used linguistic and textual features are described below in the following subsection.

5.1.1. *Bag-of-Words and N-Gram Features*

By extracting important facts from Information Retrieval (IR), various researchers have applied the Bag-of-words approach for text categorization (Feng, Banerjee et al., 2012; Li et al., 2013; Karami and Zhou, 2015). Bag-of-words forms a vector of words for a given sentence or document. By counting the occurrence frequency of each word in a review text using term frequency-inverse document frequency (tf-idf), a feature vector (see Table 2) can be formed to represent the review mathematically (Al-Najada and Zhu, 2014). Using the feature vectors, we can easily compute the similarity between two reviews by using similarity measures (e.g., cosine measure). An example of reviews along with their feature vectors is presented below:

Sentence 1: iphone is a very good product with great camera quality.

Sentence 2: I like the iphone very much as its camera quality is great.

Table 2. Feature vector representation using Bag-of-words approach.

Bag-of-words	Iphone	is	very	Good	product	with	great	camera	quality	I	like	Much	as	its
Feature vector 1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Feature vector 2	1	1	1	0	0	0	1	1	1	1	1	1	1	1

A similar approach is to use word n-grams rather than single words or unigrams (Li, Ott et al., 2014) as used in the bag-of-words technique. N-gram-based methods have shown better performance than both the

bag-of-words approach and the manual inspection technique (Ott et al., 2013). Ott et al. (2011) have highlighted the importance of bigrams and trigrams along with the unigram features to classify reviews as fake or benign. In order to capture the more subtle clues of review spamming, Fusilier et al. (2015) have used character n-grams as features (rather than word n-grams (Ott et al., 2011)) to train their classifier and shown an increase in the accuracy of results. Therefore, n-gram language models are apparently good to detect deceptive reviews and are used by most of the works published so far.

5.1.2. *POS Tags and LIWC Features*

Although n-grams have shown interesting results in detecting bogus opinions, Part-Of-Speech (POS) tags can enhance accuracies when combined with unigrams (Feng, Banerjee et al., 2012) or bigrams. By assigning the POS tags to each word of a review, a POS tagger can highlight syntactic deception clues about review spamming (Lau et al., 2011). Ott et al. (2011) trained a SVM classifier on POS word features to differentiate the writing styles of benign and malignant users. According to their analysis, spammers are inclined to write imaginative reviews by using more verbs, adverbs, or pronouns, while normal users write informative reviews to share their experiences by using more nouns or adjectives. Similarly, to detect spam comments on a post, Rădulescu et al. (2014) utilized POS information to extract unigram and bigram topics by considering nouns as unigrams and AN (adjective-noun) and NN (noun-noun) pairs as bigrams. Therefore, POS tagging is considered as an important step in feature engineering process and can be combined with tf-idf to find the frequently occurring features in a review corpus (Ong et al., 2014), which can help in distinguishing deceptive information from factual information.

To get a deeper understanding of the language that spammers use, some studies have used clues from linguistics and psychology and applied LIWC (Linguistic Inquiry and Word Count) as a tool to gain a better understanding of deceptive text writing. Ott et al. (2011) demonstrated the role of LIWC features in linguistic deception detection and showed that LIWC performs best with unigrams and bigrams on the SVM classifier. In order to improve the performance of existing review-spam detection techniques, Karami and Zhou (2015) proposed new LIWC features such as punctuation marks score, affective negative or positive feelings score, score difference between words related to personal profit and general text, and so on, and showed an accuracy of more than 93%. Although most of the existing opinion-spam detection techniques are pertinent for a specific domain (like product domain or movie domain or hotel domain), LIWC and POS features are strong indicators of linguistic deception in a multi-domain environment as well (Li, Ott et al., 2014).

5.1.3. *Other Semantic and Stylistic Features*

Besides n-grams, POS tags, and LIWC features, some papers used semantic and stylistic clues to mark a review as fake or genuine. Existing studies suggest that spammers are prone to writing similar reviews and therefore overlap in their content. It becomes necessary therefore to analyze the semantic relationship between reviews of the same user and between different users. A number of studies on this, such as Lau et al. (2011), have defined semantic content overlapping to spot fake reviews with the help of WordNet lexical database. Similar work was also shown by Lai, Xu, Lau, Li and Song (2010), who applied language modeling to discover semantically equal terms. Sandulescu and Ester (2015) extended the ordinary word similarity measure (e.g., cosine similarity) to incorporate the semantic context into it and defined a new semantic similarity measure to identify SR spammers. In order to evaluate review importance, Wang, Yan et al. (2015) computed the semantic similarity between a review and the related

object (like product, news article etc.) and ranked the reviews accordingly. Rather than applying any semantic measure, Yang (2015) demonstrated the power of store-sentiment word pairs, which are useful in capturing the deception in a review toward a specific store, and thus defined some coherence measures to find the connection between stores and words used by spammers. To capture the malicious advertising behavior of spammers, some stylistic features were proposed by Lai, Xu, Lau, Li and Jing (2010), which included percentage of capitalized words, percentage of repeated words, percentage of emotional words, percentage of personal pronouns, frequency of passive voices, and so on. Although spammers try to mimic legitimate reviewers, they lack the potential of real users. By employing readability as a measurement for writing style (Harris, 2012), authors proposed a low readability index for fake reviews than for genuine ones. They found that honest reviewers write descriptive reviews, whereas spammers write generic reviews with the intention of being able to reuse their text. By looking at the problem from a different angle, Lappas (2012) revealed the crucial factors that a spammer could consider before writing a fake review. His work considers authenticity and impact as the main ingredients a spammer should focus on to avoid creating suspicion in the customer's mind and provides different measures (e.g., stealth, coherence, readability) to evaluate these parameters effectively. To understand the strategies used by spammers before executing their plans, a similar work was presented by Banerjee and Chua (2014) in which they demonstrate two processes, collection of relevant information and articulation of such information, to write authentic-looking fake reviews.

Although linguistic features were useful to detect review spammers and have shown good accuracies, spammers are now experienced enough to write more realistic fake reviews to mimic the behavior of genuine reviewers, and therefore, investigators need to find other ways to get a set of more distinguishing features separating spam from non-spam.

5.2. Behavioral Features

Due to the difficulty involved in gathering textual evidence of review spamming, an increasing number of studies have focused on the behavioral activities of opinion spammers, which makes it easy to distinguish spammers from genuine reviewers; these methods are cost effective as well (Savage et al., 2015). Studies in the literature suggest that opinion spammers show behavior different from that of normal users, such as they praise or downgrade a particular brand of product (Jindal and Liu, 2008; Jindal et al., 2010; Shojaee et al., 2015); post the maximum number of reviews in a short span of time (Mukherjee, Kumar et al., 2013; Mukherjee, Venkataraman et al., 2013a); deviate from the majority in terms of ratings (Lim et al., 2010; Li et al., 2011; Savage et al., 2015), always provide extreme ratings (very high or very low) (Feng, Xing et al., 2012; Mukherjee, Kumar et al., 2013), repeat their ratings or reviews (Lim et al., 2010; Shojaee et al., 2015), and so on. To avoid detection, spammers write very few reviews from the same account and maintain multiple accounts to write multiple fake reviews. These fake reviews may come in bursts since most spammers probably write multiple fake reviews from different accounts over a particular time span rather than during different time intervals. To detect such suspicious spam cases, some researchers employed burst behavior analysis (Feng, Xing et al., 2012; Xie et al., 2012a; Fei et al., 2013; Mukherjee, Kumar et al., 2013) and time-sensitive behavioral features (Xie et al., 2012a; Xie et al., 2012b; Lin, Zhu, Wu et al., 2014) to indicate opinion spam and reported good accuracies. Moreover, to capture reviewers' attention, which in turn can make a good impact on product sales, spammers try to post a fake review as early as possible immediately after the product has been launched. Studies (Lim et al., 2010; Mukherjee, Kumar et al., 2013) used this feature by counting the number of reviews of a reviewer that were posted just after the launch of a product. If a reviewer is

posting most of his reviews as early as possible to make an early impact on the product's reputation, he or she may be a hired spammer. Since different studies highlight a variety of opinion spamming features of suspect opinion spammers, it is difficult to describe them all here. Therefore, to get a brief summary of spammers' behavior analysis, in Table 3, we have listed some crucial behavioral features used to spot opinion spam.

Table 3. A list of behavioral features used by state-of-the-art approaches.

Behavioral Features	References
Rating Deviation	(Jindal and Liu, 2008; Lim et al., 2010; Li et al., 2011; Mukherjee et al., 2011; Feng, Xing et al., 2012; Mukherjee et al., 2012; Fei et al., 2013; Mukherjee, Kumar et al., 2013; Mukherjee, Venkataraman et al., 2013a; Xu, 2013; Rayana and Akoglu, 2015; Savage et al., 2015; Xu and Zhang, 2015a)
Review Burst Detection	(Feng, Xing et al., 2012; Xie et al., 2012a; Xie et al., 2012b; Fei et al., 2013; Mukherjee, Kumar et al., 2013; Mukherjee, Venkataraman et al., 2013a; Lin, Zhu, Wang et al., 2014; Lin, Zhu, Wu et al., 2014; Rayana and Akoglu, 2015; Xu and Zhang, 2015a)
Maximum Number of Reviews	(Mukherjee et al., 2011; Mukherjee, Kumar et al., 2013; Mukherjee, Venkataraman et al., 2013a; Rayana and Akoglu, 2015)
Content Similarity	(Lim et al., 2010; Li et al., 2011; Mukherjee et al., 2011; Fei et al., 2013; Mukherjee, Kumar et al., 2013; Mukherjee, Venkataraman et al., 2013a; Xu, 2013; Lin, Zhu, Wang et al., 2014; Lin, Zhu, Wu et al., 2014; Rayana and Akoglu, 2015; Xu and Zhang, 2015a)
Extreme Ratings	(Feng, Xing et al., 2012; Mukherjee, Kumar et al., 2013; Rayana and Akoglu, 2015)
Early Time Frame	(Lim et al., 2010; Mukherjee et al., 2011; Mukherjee et al., 2012; Mukherjee, Kumar et al., 2013; Xu, 2013; Rayana and Akoglu, 2015)
Rating Repetition	(Lim et al., 2010; Mukherjee, Kumar et al., 2013; Lin, Zhu, Wang et al., 2014; Lin, Zhu, Wu et al., 2014; Shojaee et al., 2015)
First Review Ratio	(Jindal and Liu, 2008; Li et al., 2011; Mukherjee, Kumar et al., 2013)
Activity Time Window	(Mukherjee et al., 2011; Mukherjee et al., 2012; Xu, 2013)

Table 3 clearly shows the rating deviation from majority opinions, a sudden concentration in the number of reviews in short time spans, and content similarity among reviews of a product or a user, which have been utilized in most of the studies so far and best describe the spamming behavior of online reviewers. Moreover, some recent works (Mukherjee, Venkataraman et al., 2013a; Mukherjee, Venkataraman et al., 2013b) have shown that behavioral spamming features showed better results than linguistic features in classifying a review as fake or genuine and reduced the computational complexity considerably since they do not require working with any NLP or text mining algorithm.

5.3. Relational Features

It is sometimes easy for spammers to imitate linguistic and behavioral patterns but very hard to mimic the network structure of genuine reviewers (Ye and Akoglu, 2015). As spammers behave in a similar

manner and interact with other spammers to achieve their motives, we can use their interactions with products or other reviewers to spot malicious intent. A number of algorithms have been developed for this purpose incorporating the intricate relationships among reviews, reviewers, and products into a graphical structure. Simple reviewer behavioral attributes provide less information about spamming (Wang et al., 2012) and may lead to a high false-positive rate. Therefore, to model the more subtle behavior of spammers, Wang et al. (2011) exploited the influence of three nodes, reviews, reviewers, and products, on one another in a tri-partite graphical structure. Akoglu et al. (2013) proposed a general framework describing the association between reviewers and products in a bi-partite fashion by considering the links labeled with the review sentiment. Li, Chen et al. (2014) also used a tri-partite graphical model describing the relationship among reviews, reviewers, and IP addresses to spot anomalous behavior of review spammers. To highlight collusive spamming behavior, structural spam indicators were utilized effectively (Wang, Hou et al., 2015). They include spam group tightness, reviewer correlation in a spam group, number of products being targeted by a spam group, product reviewer ratio in a spam group, and the size of a spam group. All these features are network specific and can be easily extracted from a product-review bipartite graph. Another attempt was made by Liang et al. (2014), who calculated the spamming scores of reviewers based on reviewer-specific spamming features and boosted/hindered their scores according to supportive/conflicting ratings provided by them on different products in a product-review graph. Despite the importance of relational features in identifying opinion spam, only limited works have utilized them. These features play an important role, especially in collusive opinion spamming, where these are the only possible indicators to spot highly suspicious spam groups.

So far, a variety of spamming features have been used in diverse techniques to differentiate spam from non-spam. Linguistic, behavioral, and relational features have all been exploited by investigators, but only a few efforts (Rayana and Akoglu, 2015) have accomplished examining the combined power of these features. As it can be tricky to identify spammers by concentrating only on language or activities, it may be more efficient to use a hybrid set of features to successfully detect malicious behavior of online reviewers.

6. Categorization Based on Opinion-Spam Detection Methods

Most of the previous works have treated fake-review detection as a classification problem in which the class may be spam or non-spam for reviews, spammer or genuine for reviewers, and low- or high-quality for products. Data mining, Text mining, Machine learning, and Natural Language Processing (NLP) are the different fields that contributed to spotting fraudulent reviews or fake reviewers. As spotting fake reviews looks like a classification task, machine learning has been extensively used by researchers to detect them. Therefore, in this paper, we have demonstrated only the different machine-learning methods used by state-of-the-art approaches to identify opinion spam. Machine-learning techniques can be broadly categorized into supervised, unsupervised, and semi-supervised approaches.

6.1. Supervised Learning

Supervised learning is one that requires a labeled set of training examples to guide a model and it can be further used to predict the unknown class of new samples. As opinion spam detection is similar to a classification problem with probably two classes, spam and not spam, in the initial stages, researchers employed supervised learning to deal with the problem of fake review detection. The works (Jindal and Liu, 2007a; Jindal and Liu, 2007b; Jindal and Liu, 2008) on opinion spam detection use logistic

regression to train a supervised model by using duplicate reviews as spam class and report an Area Under Receiver Operating Characteristic (ROC) Curve (AUC) of 78% for untruthful spam reviews, which was quite good for the initial stage. The authors also showed the classification results for other kinds of spam reviews such as reviews-on-brands-only or non-reviews, and reported an AUC score of 98.7%. They then concluded that it was difficult to filter out untruthful review spam when compared to non-reviews or reviews-on-brands-only. Liu et al. (2007) trained an SVM classifier to distinguish between high- and low-quality reviews. Although review-quality assessment and review-spam identification are two different tasks, their ultimate target is to maintain the reputation of opinion-wearing sites. A low-quality review can be considered a spam but spammers also write high-quality reviews to sound authentic, and such high-quality reviews written by suspicious persons are termed "untruthful" reviews. Koven et al. (2014) focused on a different angle of opinion spam research. Instead of identifying spam reviews, they identified distinguishable features (e.g., reviewer's average star rating, review topic distribution, reviewer's personal qualities, such as rating similarity, relationship with other reviewers, total geographical location traveled, etc.) which made the review useful for readers. Some works utilized the power of supervised learning in evaluating their unsupervised models. Fei et al. (2013) employed a supervised evaluation method to assess their proposed Loopy Belief Propagation (LBP) model for spotting fraudsters and achieved an accuracy of 77.6 % when prior information (such as Amazon Verified Purchase Ratio) and local spamming behavior features were combined with that model. To evaluate different supervised learning algorithms, Banerjee et al. (2015) used a number of supervised algorithms on their proposed linguistic features. They found that among all the supervised models (e.g., SVM, Logistic Regression (LR), Naïve Bayes (NB), SVM with linear kernel, etc.), logistic regression gave the best results with an accuracy of 71.67%.

The main hurdle in supervised learning is obtaining reliable ground-truth labels for model training. It may be easy to acquire the labels for the negative class (non-fake or honest reviews) because some sites like Yelp[7] or Dianping[8] filter out fake reviews and display only honest reviews. Some studies (Mukherjee, Venkataraman et al., 2013b; Luca and Zervas, 2016) have tried to identify Yelp's filtering mechanism by considering a list of features and validating their results. Therefore, one can treat Yelp's filtered reviews as negative class examples, but a major difficulty lies in obtaining the training data for the positive class (fake reviews). Most of the works manually create the labels by examining each review individually. According to our reviewed literature, the three possible approaches for acquiring labels are as follows:

i) Some websites, such as consumerist[9], wikihow[10], or fakespot,[11] offer a way to spot fake reviews. Therefore, to obtain labels, researchers appoint human annotators to create labels (fake or not-fake) manually according to the instructions provided on these websites (Li et al., 2011).

ii) Some studies used AMT crowdsourcing tool to generate fake reviews (Ott et al., 2011; Feng, Banerjee et al., 2012; Harris, 2012; Li, Ott et al., 2014) and treated Yelp's or TripAdvisor's recommended reviews as a genuine one to train their models.

[7] http://www.yelp.com
[8] http://www.dianping.com
[9] https://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews
[10] http://www.wikihow.com/Spot-a-Fake-Review-on-Amazon
[11] http://fakespot.com

iii) Sun et al. (2013) employed a different approach to generate fake reviews. They used a synthesis approach in which a synthesizer automatically generates a fake review by replacing the sentences of a truthful review with other similar sentences in the review dataset.

Although a vast amount of research work has been done in the last decade to spot fake online reviews, only a few of them have shown promising results. The main cause for failure of most of the supervised spam detection methods is the class imbalance problem (Al-Najada and Zhu, 2014). Class imbalance arises when the data is inclined toward one specific class instead of being equally distributed among all classes. In the case of a review dataset, most of the reviews are genuine (majority class) while only a fraction of them are spam (minority class), which forces the classifier to give biased results. Therefore, it becomes difficult for a classifier to classify the spam and non-spam reviews accurately when proper care is not given while training the model with imbalanced data. Al-Najada and Zhu (2014) proposed a solution to deal with imbalanced review data in which random sampling is used to produce a number of balanced datasets, learn about a classifier on each dataset, and finally use an ensemble of methods to label a review (fake or non-fake) by a majority voting of classifiers' results.

In short, supervised learning provides an excellent way to spot spam opinions but its own limitations sometimes cause it to be stuck at the very first stage. Although some works have shown good results using supervised techniques (Ott et al., 2011; Feng, Banerjee et al., 2012), their results are domain specific. For example, a model trained on a product domain may not produce accurate results in the hotel or movie domain. Therefore, the need arises to scrutinize supervised learning further and propose a general framework that can be applied in any review domain effortlessly.

6.2. Unsupervised Learning

Another possible way is to employ unsupervised learning techniques capable of monitoring the behavior of reviewers and raising an alarm whenever abnormal behavior occurs. As we stated earlier, supervised learning is applicable only when enough ground-truth labels are available for model training. However, it is often difficult to obtain such labels, especially in cases of review datasets where spammers are so clever in posting reviews that even human experts find it difficult to label them. This makes supervised learning obsolete for review spam detection. As a result, researchers proposed unsupervised approaches as an alternative to spot suspicious reviewer behavior. Jindal et al. (2010) proposed a generic model using class association rule mining to spot the anomalous rating behavior of reviewers. By analyzing the ratings given by a customer to different products of a brand, they constructed important rules showing abnormal reviewer behaviors and concluded that reviewers who are always positive about a brand of products remarking on no downsides or vice versa are likely to be indulging in spam activities. More importantly, their framework was domain independent and therefore applicable to other domains as well, which was not so with supervised learning. Moreover, spammers usually reproduce their reviews and alter a few conceptual terms to sound like a genuine user. Therefore, some studies used inferential language modeling techniques using content overlapping and concept association between reviews to spot spam opinions in an unsupervised manner (Lai, Xu, Lau, Li and Jing, 2010; Lai, Xu, Lau, Li and Song, 2010; Lau et al., 2011). Earlier researchers were more directed toward the textual clues responsible for review spamming, while current work focuses on the network structure of reviewers and products to get a clearer picture of opinion spamming. This was computationally efficient, as it did not involve any kind of NLP or text mining algorithms, which require significant amount of cost.

According to social learning theory in criminology (Akers, 2011), the physical and social environment play a critical role in determining how criminals learn a particular behavior. Therefore, people around a criminal can provide clues to spot the criminal's activities. By the same yardstick, Wang et al. (2011) exploited the influence of neighboring nodes on a node (e.g., review, store, or reviewer) in the review network and demonstrated how the neighbors (reviewers on the same product or store) of a node can disclose clues of opinion spamming. They achieved a precision of 49% through the human evaluation method and justified their method, as they were the first to detect complex spam activities, which had been almost neglected by previous works. Since then, a number of graph-based unsupervised algorithms (Akoglu et al., 2013; Peng, 2013; Rayana and Akoglu, 2015) have been proposed to deal with the problem of detecting opinion spammers. Furthermore, a few works have utilized network-specific spamming clues to detect a cluster of spammers in a purely unsupervised manner. Wang, Hou et al. (2015) suggested an alternative method to the FIM (Frequent Itemset Mining), namely, a bipartite graph projection, to find a loosely connected group of spammers in a convenient way, which is usually the case in real-life networks. A loose spammer group is a group of reviewers in which not every reviewer has reviewed each product in the group; instead, a reviewer would review all or only a portion of the products, with some common products reviewed by other members of the group. Their results show a good precision (approx. more than 0.8 for top 100 and 200 lists) and the method they proposed was able to identify more subtle clusters of spammers when FIM-based methods became obsolete. Another study (Ye and Akoglu, 2015) argued that spammers distort the network structure of real users by invoking some unusual patterns into the original network; therefore, they proposed an unsupervised framework to examine in depth the network traits of opinion spammer groups. To quantify the statistical distortions originating from spammer behavior, they computed the degree and pagerank centralities of different products at the local and global level, respectively, and devised a clustering algorithm, GroupStainer, to obtain a hierarchy of clusters representing suspicious spamming groups.

6.3. Semi-supervised Learning

Although supervised learning proved to be useful in detecting opinion spam, it required a huge amount of manual effort and cost too many resources to obtain a sufficient amount of labeled data. As a result, researchers began to employ semi-supervised learning to lessen the manual efforts involved in supervised learning. Semi-supervised supervised learning utilizes the labeled data (only a small amount) to label the huge amount of unlabeled data and uses both types of data to predict the class of a sample. Usually in case of a fake review data, a huge number of reviews are unlabeled, which makes annotation a tedious task and therefore semi-supervision may be the best choice to cope with the problem of manual labeling. A likely semi-supervised approach is to use a co-training algorithm to train two separate views of training data by a classifier. Li et al. (2013) employed such a co-training framework using review features and reviewer-specific features in the first and second view, respectively, to annotate a huge amount of unlabeled data by using only a small training set of annotated reviews. In this manner, they finally obtained enough reviews having labels as fake, which were in agreement with both the views of their classifier. Some other works have shown that PU (Positive-Unlabeled) learning is a good alternative to supervised learning when enough ground-truth labels are not available. In classifying an object, whenever the target class contains only few instances of known labels and most others belong to the unknown set (i.e., unlabeled set), then PU learning has the capacity to differentiate the positive and negative instances from the unlabeled set of objects. Li, Chen et al. (2014) employed the PU learning framework to improve the performance of their supervised classifier. They developed two models to collectively classify a

review, IP address, and reviewer as genuine or suspicious. The first model assumes unlabeled instances as the negative ones and applies supervised learning for classification, while the other model employs PU learning to update the initial labels of nodes from the unlabeled set by leveraging neighboring label information in the multipartite network structure of reviews, reviewers, and IPs. The F1-score they achieved shows the effectiveness of the PU learning framework over the supervised one even with small-sized training datasets. The work of Montes-y-Gómez and Rosso (2013) shows how PU learning outperforms other supervised classification approaches (e.g., Naïve Bayes or SVM) with their lack of negative class examples. They iteratively find the positive class tuples from the unlabeled set and retrain a classifier after eliminating these positive tuples. They achieved a f-measure of 0.837, which is comparable with the work of Ott et al. (2011), especially when only 25% of the deceptive opinion data was used for training. Further, in a subsequent paper, Fusilier et al. (2014) take both the deceptive opinions' polarities, positive and negative, into account and demonstrate the power of PU learning in identifying both types of deceptive opinion spam. Their results indicate an F1-score of approximately 0.8 and 0.7 for positive and negative opinion spam, respectively, which clearly depicts how difficult it is to detect negative opinion spam automatically when compared to positive opinion spam.

To gain a better understanding of the diverse techniques used to detect opinion spam, we have summarized a few state-of-the-art published works in Table 4. Table 4 is self-explanatory; it includes the different machine-learning techniques used by previous works along with their results, spamming features, detection targets, datasets, and the publication year. This table provides a quick view of the research work done so far on the problem of opinion spam detection along with the results.

Table 4. Summary of existing state-of-the-art approaches.

Ref.	Machine Learning Technique	Detection Target	Spamming Features	Results/Accuracies	Dataset Used	Publication Year
Jindal and Liu (2008)	LR Classifier	Fake Reviews	Review specific and reviewer specific	AUC-78%	Amazon dataset- 5.8 million reviews and 2.14 million reviewers.	2008
Lim et al. (2010)	Unsupervised	Product Review Spammers	Rating deviation, Rating / text repetition	Cohen's kappa values 0.48 and 0.64 showing moderate human agreement	Amazon MProducts dataset- 313,120 reviewers, 32,075 products and 404,637 reviews.	2010
Lau et al. (2011)	Unsupervised Language Modeling Approach	Fake Reviews	Content overlapping, Semantic term association	Kappa K=0.81 (almost perfect agreement), both logistic average misclassification rate (lam%) = 1.70 and area above the ROC curve (1-AUC)= 0.1346 show the lowest error rate	Amazon dataset- 2,318,989 reviews from 10 product categories.	2011

Wang et al. (2012)	Unsupervised graphical approach	Store Review spammers	Relationship among honesty of reviews, trustiness of reviewers and reliability of stores	Cohen's kappa k= 0.603 (substantial agreement among human evaluators), Precision = 49%	Resellerraings.com dataset- 343,603 reviewers, 408,470 reviews and 14,561 stores.	2012
Ott et al. (2011)	SVM and NB classifiers	Opinion Spam	LIWC features, POS tags, Unigrams, Bigrams and Trigrams features	Best result: Accuracy=89.8% when LIWC and Bigram features are applied on SVM	Hotel review dataset- 400 genuine reviews from TripAdvisor and 400 deceptive reviews from AMT.	2011
Li et al. (2011)	Semi-supervised co-training framework	Fake Reviews	Unigrams, Bigrams, Sentiment features, Product specific features, Reviewer Profile features, behavior features	Best results: Precision= 0.641 Recall= 0.621 F-Score= 0.631	Epinions dataset- 60k reviews After human annotation, dataset reduced to 6000 reviews with 1398 spam reviews.	2011
Mukherjee et al. (2012)	Unsupervised relational model	Opinion spammer groups	Behavior features: Group Time Window, Group content similarity, Group rating deviation, Group size ratio, Group early time frame	Kappa k=0.79 (close to perfect agreement), supervised evaluation shows an AUC= 0.95 and 0.93 for different thresholds.	Amazon product review dataset- 109, 518 reviews, 53,469 reviewers and 39,392 products.	2012
Feng, Banerjee et al. (2012)	SVM classifier	Fake Reviews	Unigrams, bigrams, Pos tags, PCFG (Probabilistic Context Free Grammar) production rules	Best result: Accuracy= 91.2% on combined unigrams and PCFG features	TripAdvisor hotel domain, Yelp restaurant domain and AMT generated essay domain datasets.	2012
Mukherjee, Kumar et al. (2013)	Unsupervised Bayesian framework	Product review spammers	Behavior features: Content similarity, Review burstiness, First review ratio, Extreme ratings, Rating deviation, Early time frame	Human evaluation shows kappa k=0.73 (substantial agreement), Evaluation by classification shows great improvement over the existing approaches	Amazon dataset: 50,704 reviewers, 985,765 reviews and 112,055 products.	2013
Li, Chen et al. (2014)	Semi-supervised PU learning framework	Fake Reviews	Unigram and Bigram features, Behavioral features of reviewers and IPs, Relational	Collective Positive Unlabeled learning performs better in terms of recall, precision, accuracy and f1-score	Dianping restaurant dataset: 9765 reviews of 500 restaurants from 9067 users and 5535 IPs.	2014

			features of reviewers, reviews and IPs.	measures.		
Rayana and Akoglu (2015)	Unsupervised framework SPEAGLE and Semi-supervised version SPEAGLE+	Fake Reviews, Review Spammers and Targeted Spam Products	Review content and meta-data features, Behavioral features, Relational features	In SPEAGLE, AUC= 0.6905 for users and 0.7887 for reviews ranking. In SPEAGLE+, AUC= 0.7078 for users and 0.7951 for reviews ranking	Three versions of Yelp restaurant dataset: YelpChi, YelpNYC, YelpZip.	2015
Ye and Akoglu (2015)	Unsupervised graphical approach	Opinion spammer groups	Network specific features: Neighbor diversity of a product (degree and page rank centrality), Self similarity with the entire network	AUC of precision-recall curve shows the effectiveness of method to capture highly suspicious spammer groups	iTunes App review dataset: 966,808 users, 15,093 products and 1,132,329 connections. Amazon dataset: 2,146,074 users, 1,230,916 products and 5,838,061 connections.	2015

7. Results and Summary

After collecting and carefully studying a number of superior quality research papers from reputed journals and conferences, we have summarized opinion spam detection in terms of per-year distribution of research articles, detection targets, spamming features, and detection methods employed by the different studies so far. In this section, we sum up the paper and present a few interesting results pointing to future research gaps to be filled by new researchers. Research on opinion spam detection has gained attention since 2007 and evolved on a large scale due to the increasing focus of both academicians and researchers. Figure 5 shows the per-year distribution of research articles on opinion spam detection from 2007 to 2015. The numeric quantity on each bar represents the number of research articles published in a year, where the year is shown on the X-axis. It is quite clear from the figure that there is an increased concentration of research articles on opinion spam detection in the last four years with researchers showing more interest in this field. Compared to other types of spam, opinion spam is now widespread and difficult to resolve with many studies focused on making some progress in the field. The problem continues to remain unresolved with hopes for more efficient methods to achieve a more satisfactory solution in the near future.

Furthermore as per our collected literature, a number of articles have been published in conference publications. The percentage of articles published in conferences is far higher than in journal publications as can be seen from Fig. 6. Figure 6 evidently shows that researchers were more directed toward conference publication, which includes 65% of the total research articles that appeared so far. On the contrary, an extremely small fraction (13%) of the work has been published in reputed journals. This indicates a clear research gap for the practitioners and raises the question, “Why are there fewer articles on this topic in journals as compared to those in conferences?” To make sound progress on a specific research topic, a sufficient amount of work should be published in journal publications.

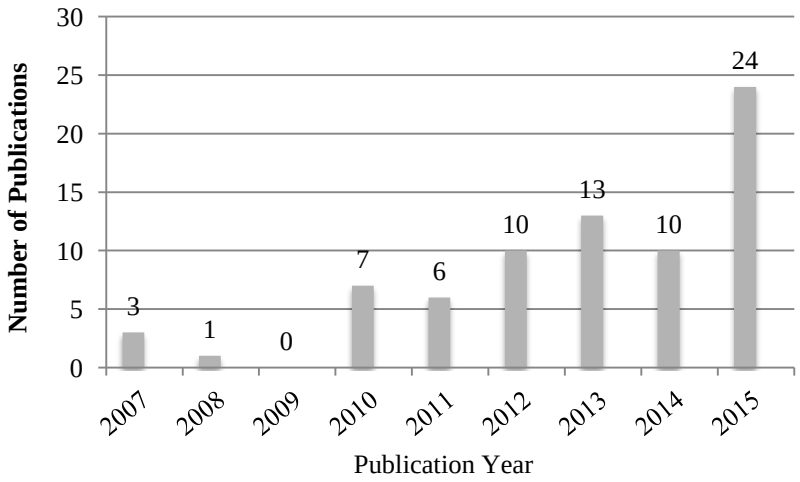

Fig. 5. Per year distribution of research articles on opinion spam detection.

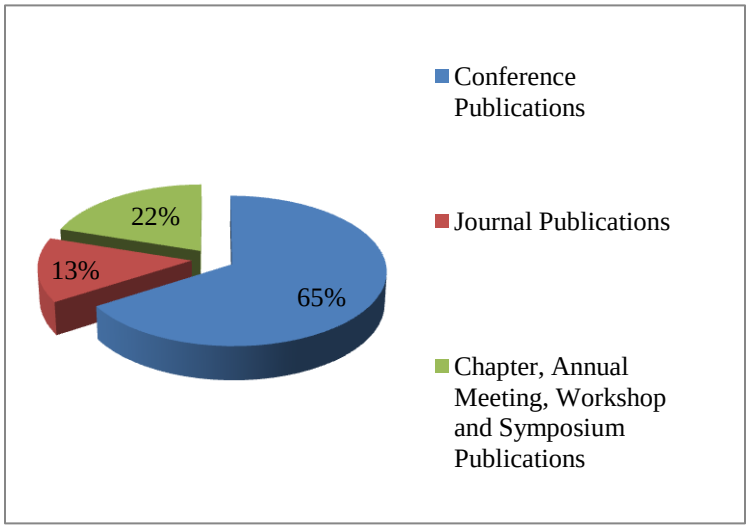

Fig. 6. Distribution of research articles according to publications.

As only a few works have approached some reputed journals, there still exists a need for novel and innovative opinion-spam detection techniques capable of capturing the behavior of opinion spammers as early as possible.

As mentioned in section 4, researchers have three different detection targets to spot the fraudulent activities of reviewers resulting in opinion spamming. The percentage of research articles focusing on different detection targets is shown in Fig. 7. Each sector represents the percentage of articles focusing on a particular target. As per this figure, an enormous amount of work has been done to detect opinion spam while only a few efforts were made in the direction of individual or group spammer detection. Moreover, only a few works have explored the detection of opinion spam and spammers in an integrated framework. As spammers are the primary source of opinion spamming, it is of utmost importance to investigate their suspicious behavior. However, unfortunately, this has not been given sufficient attention by most of the

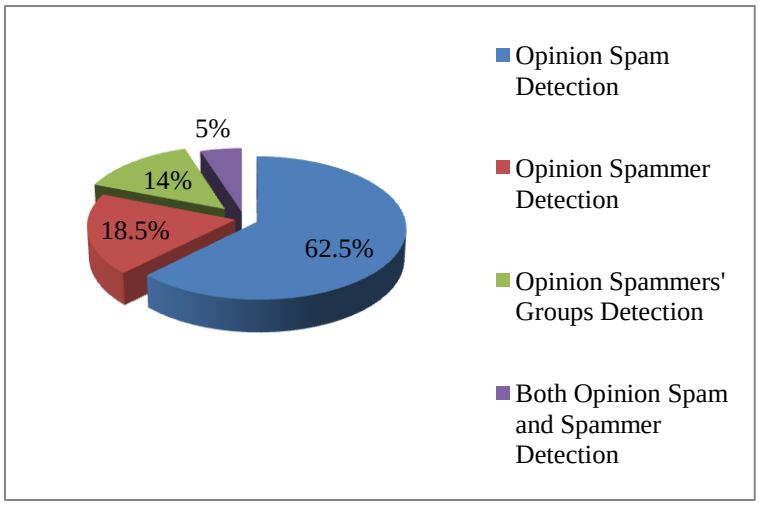

Fig. 7. Distribution of research articles according to detection targets.

previous works so far. Spammers' abnormal behavior can be used to detect them and similar reviewers in an effective manner. Future research should consider this aspect when attempting to detect fake reviews and malicious reviewers at first glance.

Moreover, recent literature reveals the correlation among spammers working on a single product or group of products to maximize their fake impact, but the study lacks the methods (only 14% success as shown in Fig. 7) to deal with the collusive behavior of spammers. More general and versatile methods need to be explored to cope with the problem of fake review detection, especially in the case of group spamming.

A brief summary of spamming features used to detect opinion spam is presented in Fig. 8. Although it is a little complicated to spot untruthful forged reviews by merely considering a writer's writing style and ignoring the other parameters of spamming, the maximum number of studies (40% of our collected literature) used linguistic and textual clues of lying as illustrated in Fig. 8. This is possible only because of the availability of a large number of NLP and text mining techniques.

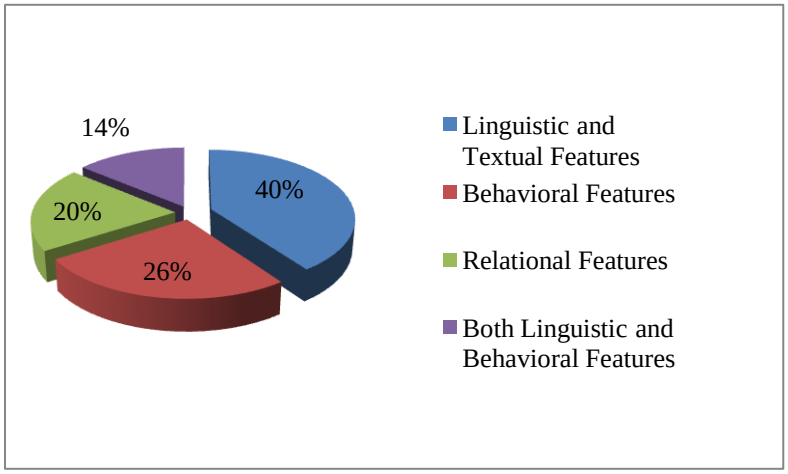

Fig. 8. Distribution of research articles according to spamming features.

Behavior and relational features are almost equally employed by previous studies. Another important observation is that all these features are used independently in the literature. Therefore, a possible future research direction is to create an optimal set of features from each category to detect review spam cases with good accuracy.

Furthermore, different studies employed different machine-learning approaches to deal with opinion spam detection as discussed in section 6. In Figure 9, we see the distribution of machine-learning techniques used in the existing literature to detect opinion spam. Each sector in this figure represents the percentage of research papers that employed a particular machine-learning technique. According to our reviewed literature on opinion spam, it is clear from this figure that both supervised and unsupervised learning have received equal attention from researchers in the past. One of the major hurdles in detection of spam reviews is obtaining scalable ground-truth datasets for model training. Semi-supervised learning has the ability to grow the size of the training data sufficiently and successfully. However, despite the significance of semi-supervised learning, only a few researchers have tried (only 8% of the total reviewed literature) to detect fake reviews in the semi-supervised fashion as shown in Fig. 9. This clearly indicates a research gap for future researchers. We need to further investigate various semi-supervised learning methods to cease the abusive actions of opinion spammers at the early stage of opinion spamming. Consequently, online consumers and the industry can benefit from the high level of trust and better quality product management, respectively.

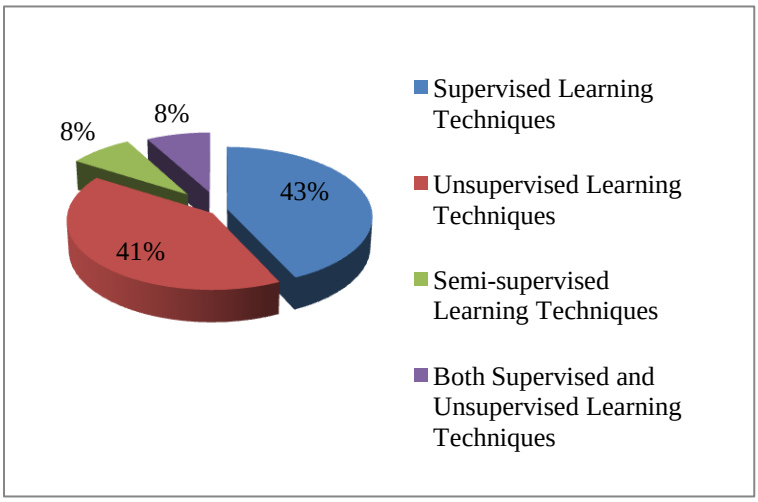

Fig. 9. Distribution of research articles according to machine-learning techniques.

8. Conclusion

Due to the increasing dependence of consumers on online reviews, fraudsters are flooding the review system by writing fake opinions on targeted products or organizations. Consumer trust on the opinion-wearing websites is therefore declining. In this study, we have focused on the problem of opinion spam detection, providing a brief and significant impression of related research work carried out in the last decade. By looking at the problem from different angles, we categorized the available literature on fake review detection according to three different parameters: detection targets, spamming features, and techniques employed by previous works. To understand the current progress on opinion-spam detection research, we briefly described some of the crucial spamming features and methods adopted in the existing studies along with their accuracies. We also presented some valuable results indicating future research

directions for new researchers and practitioners to fill the gaps. The results shown in this paper are as follows:

i) A major challenge in detection fake reviews is to come across with the best set of features that best describes the spamming behavior of reviewers. As the number of available features is very limited, many existing approaches suffer in terms of accuracy. One of the results indicates that most of the previous works on opinion spam detection used the linguistic/textual and behavioral features individually and that only a few works relied on the combined power of these features. Therefore, to boost the performance of existing opinion spam filters, an optimal set of hybrid features need to be provided to the model for training and so better prediction.
ii) Again, one of the biggest challenges in the identification of fake opinions is to obtain a huge and reliable ground-truth dataset to train a machine-learning model. According to our observations, semi-supervised learning has not received the due attention of researchers in detecting fake reviews (only 8% of the reviewed literature as shown in Fig. 9). Therefore, a possible future research direction is to explore a variety of semi-supervised approaches to resolve the problem of ground-truth generation, which will in turn help in effective detection of fake reviews.
iii) As per our reviewed literature, detection of collusive opinion spammers groups is a neglected part of opinion-spam detection research. Only a few recent studies have highlighted the collusive behavior of opinion spammers, but we still lack suitable methods with good accuracy. In recent times, since spammers act in groups to make their impact as high as possible, all network-specific characteristics can play a major role in spotting their correlated malicious behavior.

References

Akers, RL (2011). Social learning and social structure: A general theory of crime and deviance. Transaction Publishers.

Akoglu, L, R Chandy and C Faloutsos (2013). Opinion fraud detection in online reviews by network effects. In *Proceedings of the 7th AAAI International Conference on Weblogs and Social Media* (ICWSM'13), AAAI, pp. 2-11.

Algur, SP, AP Patil, PS Hiremath and S Shivashan (2010). Conceptual level similarity measure based review spam detection. In *Proceedings of the 2010 International Conference on Signal and Image Processing* (ICSIP), IEEE, pp. 416-423.

Allahbakhsh, M and A Ignjatovic (2015). An iterative method for calculating robust rating scores. IEEE *Transactions on Parallel and Distributed Systems*, 26(2), 340-350.

Al-Najada, H and X Zhu (2014). iSRD: Spam review detection with imbalanced data distributions. In *Proceedings of the 15th IEEE international conference on Information Reuse and Integration* (IRI), IEEE, pp. 553-560.

Banerjee, S and AYK Chua (2014). Understanding the process of writing fake online reviews. In *Proceedings of the 9th International Conference on Digital Information Management* (ICDIM), IEEE, pp. 68-73.

Banerjee, S, AYK Chua and JJ Kim (2015). Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, ACM, p. 88.

Chakraborty, M, S Pal, R Pramanik and CR Chowdary (2016). Recent developments in social spam detection and combating techniques: A survey. *Information Processing and Management*, *52*(6), 1053-1073.

Chen, C, K Wu, V Srinivasan and X Zhang (2015). A comprehensive analysis of detection of online paid posters. In *Recommendation and Search in Social Networks*, Springer International Publishing, pp. 101-118.

Chen, YR and HH Chen (2015). Opinion spam detection in web forum: a real case study. In *Proceedings of the 24th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 173-183.

Chengzhang, J and DK Kang (2015). Detecting spamming stores by analyzing their suspicious behaviors. In *Proceedings of the 2015 17th International Conference on Advanced Communication Technology* (ICACT), IEEE, pp. 502-507.

Choo, E, T Yu and M Chi (2015). Detecting opinion spammer groups through community discovery and sentiment analysis. In *Data and Applications Security and Privacy* XXIX, Springer International Publishing, pp. 170-187.

Crawford, M, TM Khoshgoftaar, JD Prusa, AN Richter and H Al-Najada (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, 2(1), 1-24.

Delany, SJ, M Buckley and D Greene (2012). SMS spam filtering: methods and data. *Expert Systems with Applications*, 39(10), 9899-9908.

Fei, G, A Mukherjee, B Liu, M Hsu, M Castellanos and R Ghosh (2013). Exploiting burstiness in reviews for review spammer detection. In *Proceedings of the 7th AAAI International Conference on Weblogs and Social Media* (ICWSM'13), AAAI, pp. 175-184.

Feng, S, L Xing, A Gogar and Y Choi (2012). Distributional Footprints of Deceptive Product Reviews. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media* (ICWSM), AAAI, pp. 98-105.

Feng, S, R Banerjee and Y Choi (2012). Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2, 171-175.

Fusilier, DH, M Montes-y-Gómez, P Rosso and RG Cabrera (2014). Detecting positive and negative deceptive opinions using PU-learning. *Information Processing and Management*, 51(4), 433-443.

Fusilier, DH, M Montes-y-Gómez, P Rosso and RG Cabrera (2015). Detection of opinion spam with character n-grams. In *Computational Linguistics and Intelligent Text Processing*, Springer International Publishing, pp. 285-294.

Guzella, TS and WM Caminhas (2009). A review of machine learning approaches to spam filtering. *Expert Systems with Applications*, 36(7), 10206-10222.

Harris, C (2012). Detecting deceptive opinion spam using human computation. In *Workshops at AAAI on Artificial Intelligence*, pp. 87-93.

Heydari, A, M ali Tavakoli, N Salim and Z Heydari (2015). Detection of review spam: A survey. *Expert Systems with Applications*, 42(7), 3634-3642.

Hu, N, L Liu, and V Sambamurthy (2011). Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3), 614-626.

Jindal, N and B Liu (2007b, October). Analyzing and detecting review spam. In *Proceedings of the 7th IEEE International Conference on Data Mining*, IEEE, pp. 547-552.

Jindal, N and B Liu (2008). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, ACM, pp. 219-230.

Jindal, N, and B Liu (2007a, May). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, ACM, pp. 1189-1190.

Jindal, N, B Liu and EP Lim (2010). Finding unusual review patterns using unexpected rules. In *Proceedings of the 19th ACM international conference on Information and Knowledge Management*, ACM, pp.1549-1552.

Karami, A and B Zhou (2015). Online review spam detection by new linguistic features. In *proceedings of the 2015 iConference*.

Koven, J, H Siadati and CY Lin (2014). Finding valuable yelp comments by personality, content, geo, and anomaly analysis. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop* (ICDMW), IEEE, pp. 1215-1218.

Lai, CL, KQ Xu, RY Lau, Y Li and D Song (2010). High-order concept associations mining and inferential language modeling for online review spam detection. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshop* (ICDMW), IEEE, pp. 1120-1127.

Lai, CL, KQ Xu, RY Lau, Y Li and L Jing (2010). Toward a language modeling approach for consumer review spam detection. In *Proceedings of the 7th IEEE International Conference on e-Business Engineering* (ICEBE), IEEE, pp. 1-8.

Lappas, T (2012). Fake reviews: The malicious perspective. In *Natural Language Processing and Information Systems* (NLDB), pp. 23-34.

Lau, RY, SY Liao, RCW Kwok, K Xu, Y Xia, and Y Li (2011). Text mining and probabilistic language modeling for online review spam detecting. ACM *Transactions on Management Information Systems*, 2(4), 1-30.

Lee, SY, L Qiu and A Whinston (2015). The perils of online manipulation. In *Proceedings of the 2015 48th Hawaii International Conference on System Sciences* (HICSS), IEEE, pp. 4864-4873.

Li, F, M Huang, Y Yang and X Zhu (2011). Learning to identify review spam. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 22(3), 2488-2493.

Li, H, Z Chen, A Mukherjee, B Liu and J Shao (2015). Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In *Proceedings of the 9th AAAI International Conference on Web and Social Media*, AAAI, pp. 634-637.

Li, H, Z Chen, B Liu, X Wei and J Shao (2014). Spotting fake reviews via collective positive-unlabeled learning. In *Proceedings of the 2014 IEEE International Conference on Data Mining* (ICDM), IEEE, pp. 899-904.

Li, J, C Cardie and S Li (2013). TopicSpam: a Topic-Model based approach for spam detection. In *Proceedings of the 51st Annual Meeting of Association for Computational Linguistics*, 2, 217-221.

Li, J, M Ott, C Cardie and EH Hovy (2014). Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pp. 1566-1576.

Liang, D, X Liu and H Shen (2014). Detecting spam reviewers by combining reviewer feature and relationship. In *Proceedings of the 2014 international conference on Informative and Cybernetics for Computational Social Sciences* (ICCSS), IEEE, pp. 102-107.

Lim, EP, VA Nguyen, N Jindal, B Liu and HW Lauw (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ACM, pp. 939-948.

Lin, Y, T Zhu, H Wu, J Zhang, X Wang and A Zhou (2014). Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM), IEEE, pp. 261-264.

Lin, Y, T Zhu, X Wang, J Zhang and A Zhou (2014). Towards online review spam detection. In *Proceedings of the companion publication of the 23rd international conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 341-342.

Liu, B (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, *5*(1), 1-167.

Liu, J, Y Cao, CY Lin, Y Huang and M Zhou (2007). Low-Quality Product Review Detection in Opinion Summarization. In EMNLP-CoNLL, pp. 334-342.

Lu, Y, L Zhang, Y Xiao and Y Li (2013). Simultaneously detecting fake reviews and review spammers using factor graph model. In *Proceedings of the 5th Annual ACM Web Science Conference*, ACM, pp. 225-233.

Luca, M and G Zervas (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427.

Ma, Y and F Li (2012). Detecting review spam: Challenges and opportunities. In *Proceedings of the 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing* (CollaborateCom), IEEE, pp. 651-654.

Montes-y-Gómez, M and P Rosso (2013). Using PU-learning to detect deceptive opinion spam. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analytics* (WASSA), ACL, pp. 38-45.

Mukherjee, A, A Kumar, B Liu, J Wang, M Hsu, M Castellanos and R Ghosh (2013). Spotting opinion spammers using behavioral footprints. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, ACM, pp. 632-640.

Mukherjee, A, B Liu and N Glance (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*, ACM, pp. 191-200.

Mukherjee, A, B Liu, J Wang, N Glance and N Jindal (2011). Detecting group review spam. In *Proceedings of the 20th international conference companion on World Wide Web*, ACM, pp. 93-94.

Mukherjee, A, V Venkataraman, B Liu and N Glance (2013a). Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews. UIC-CS-03-2013 Technical Report, University of Illinois at Chicago.

Mukherjee, A, V Venkataraman, B Liu and NS Glance (2013b, July). What yelp fake review filter might be doing?. In *Proceedings of the 7th AAAI International Conference on Weblogs and Social Media* (ICWSM), AAAI, pp. 409-418.

Ong, T, M Mannino and D Gregg (2014). Linguistic characteristics of shill reviews. *Electronic Commerce Research and Applications*, 13(2), 69-78.

Ott, M, C Cardie and JT Hancock (2013, June). Negative Deceptive Opinion Spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (HLT-NAACL), pp. 497-501.

Ott, M, Y Choi, C Cardie and JT Hancock (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1, 309-319.

Peng, Q (2013). Store review spammer detection based on review relationship. In *Advances in Conceptual Modeling*, Springer International Publishing, pp. 287-298.

Radulescu, C, M Dinsoreanu and R Potolea (2014). Identification of spam comments using natural language processing techniques. In *Proceedings of the 2014 IEEE international conference on Intelligent Computer Communication and Processing* (ICCP), IEEE, pp. 29-35.

Rayana, S and L Akoglu (2015). Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 985-994.

Sandulescu, V and M Ester (2015). Detecting singleton review spammers using semantic similarity. In *Proceedings of the 24th International Conference on World Wide Web Companion*, International World Wide Web Conferences Steering Committee, pp. 971-976.

Savage, D, X Zhang, X Yu, P Chou and Q Wang (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems with Applications*, 42(22), 8650-8657.

Sharma, K and KI Lin (2013). Review spam detector with rating consistency check. In *Proceedings of the 51st ACM Southeast Conference* (ACMSE'13), ACM, p. 34.

Sheibani, AA (2012). Opinion mining and opinion spam: A literature review focusing on product reviews. In *Proceedings of the 6th International Symposium on Telecommunications* (IST), IEEE, pp. 1109-1113.

Shojaee, S, A Azman, M Murad, N Sharef and N Sulaiman (2015). A framework for fake review annotation. In *Proceedings of the 2015 17th UKSIM-AMSS International Conference on Modeling and Simulation*, IEEE, pp. 153-158.

Spirin, N and J Han (2012). Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2), 50-64.

Sun, H, A Morales and X Yan (2013). Synthetic review spamming and defense. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, ACM, pp. 1088-1096.

Wang, G, S Xie, B Liu and PS Yu (2011). Review Graph based Online Store Review Spammer Detection. In *Proceedings of the 11th IEEE International Conference on Data Mining*, IEEE, pp. 1242-1247.

Wang, G, S Xie, B Liu and PS Yu (2012). Identify online store review spammers via social review graph. ACM *Transactions on Intelligent Systems and Technology* (TIST), 3(4), 61.

Wang, JZ, Z Yan, LT Yang and BX Huang (2015). An approach to rank reviews by fusing and mining opinions based on review pertinence. *Information Fusion*, 23, 3-15.

Wang, Z, T Hou, D Song, Z Li and T Kong (2015). Detecting review spammer groups via bipartite graph projection. *The Computer Journal*, *59*(6), 861-874.

Wu, G, D Greene, B Smyth and P Cunningham (2010). Distortion as a validation criterion in the identification of suspicious reviews. In *Proceeding of the First Workshop on social Media Analytics* (SOMA), ACM, pp. 10-13.

Xie, S, G Wang, S Lin and PS Yu (2012a, August). Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, ACM, pp. 823-831.

Xie, S, G Wang, S Lin and PS Yu (2012b, April). Review spam detection via time series pattern discovery. In *Proceedings of the 21st international conference companion on World Wide Web*, ACM, pp. 635-636.

Xu, C (2013). Detecting collusive spammers in online review communities. In *Proceedings of the 6th Workshop on Ph. D. students in Information and Knowledge Management*, ACM, pp. 33-40.

Xu, C and J Zhang (2015a, June). Combating product review spam campaigns via multiple heterogeneous pairwise features. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, pp. 172-180.

Xu, C and J Zhang (2015b, November). Towards collusive fraud detection in online reviews. In *Proceedings of the 2015 15th IEEE International Conference on Data Mining* (ICDM), IEEE, pp. 1051-1056.

Xu, C, J Zhang, K Chang and C Long (2013). Uncovering collusive spammers in Chinese review websites. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management* (CIKM), ACM, pp. 979-988.

Yang, X (2015). One methodology for spam review detection based on review coherence metrics. In *Proceedings of the 2014 international conference on Intelligent Computing and Internet of Things* (ICIT), IEEE, pp. 99-102.

Ye, J and L Akoglu (2015). Discovering opinion spammer groups by network footprints. In *Machine Learning and Knowledge Discovery in Databases*, Springer International Publishing, pp. 267-282.